Project Report

Movie Analyser

spimpalk@buffalo.edu

Introduction:

The Internet Movie Database (IMDb) is a website that serves as an online database of world cinema. This website contains a large number of public data on films such as the title of the film, the year of release of the film, the genre of the film, the audience, the duration of the film, the summary of the film, actors and much more.

Objective:

To scrape the data available on the IMDb website for the movies in the year 2019-2020 and to interpret the data based on analyses and visualisation.

Libraries Used:

Beautifulsoup4, matplotlib, numpy, pandas, scikit-learn, seaborn, selenium

Data preparation:

- We have scraped data from IMDB website using the following link

https://www.imdb.com/search/title/?title_type=feature&release_date=2019-01-01,2020-01-01

- Selenium webdriver and Beautifulsoup4 are required to navigate through the web pages and scraping the data from web page respectively
- We collected Movie Names, Release Date, Number of votes, Average Rating, Movie genres from the website
- We collected all the above mentioned data for 250 different movies.
- We navigated through multiple pages to collect the required data(Eg. Release date is on a different page to the basic movie data).

Database Structure:

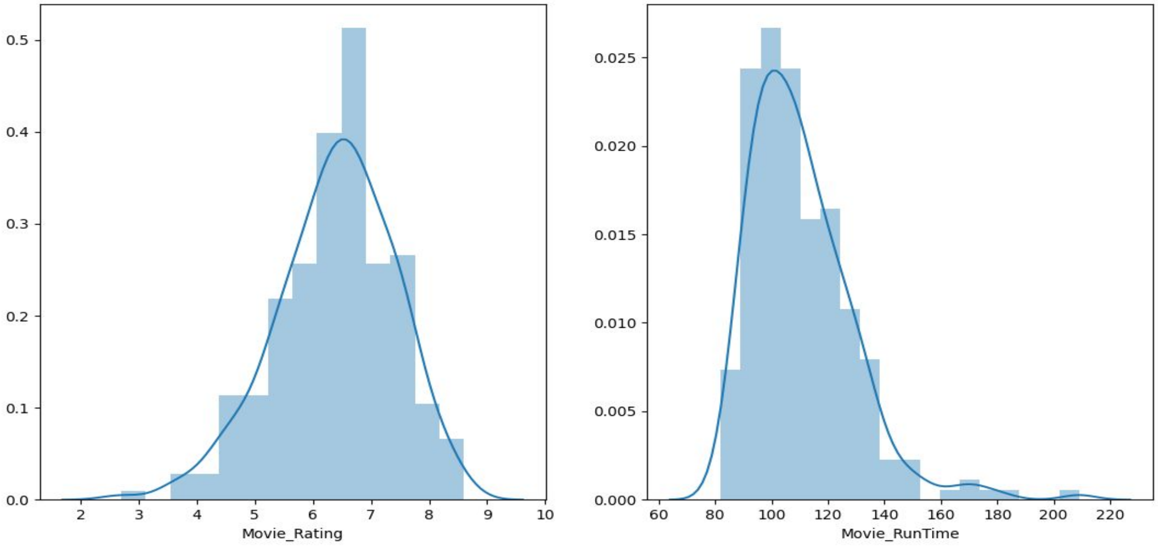
- We have created a Database called Movie_Analyzer.db
- In the database, we have created four different tables named MOVIE_MASTER,MOVIE_GENRE_MAPPER,MOVIE_GENRE_MASTER,MOVIE_RELEASE_DATE
- MOVIE_MASTER table contains the basic details of the movie. MOVIE_MASTER has four columns named Movie_Name,Movie_RunTime,Movie_Rating,Movie_Votes_Count. Movie_Name is the primary key for the table
- MOVIE_GENRE_MASTER table contains the list of genres from the movie we have scraped. Each genre contains an ID which is the primary key for the table. The table contains two columns Genre_ID(primary key) and Genre
- MOVIE_GENRE_MAPPER table maps both the Genres and Movies. One movie can have many genres (Eg. Rom-Com movie comes under both Romantic and Comedy

Genres). This table has three columns Movie_Name,Movie_Genre,ID. Movie_Name is foreign key to Movie_Name in MOVIE_MASTER table. ID is a primary key of the table

- MOVIE_RELEASE_DATE table contains the release dates of the movies. The table contains three columns named PK, Movie_Name and Release_Date. Movie_Name is foreign key to Movie_Name in MOVIE_MASTER table. PK is the primary key of the table.

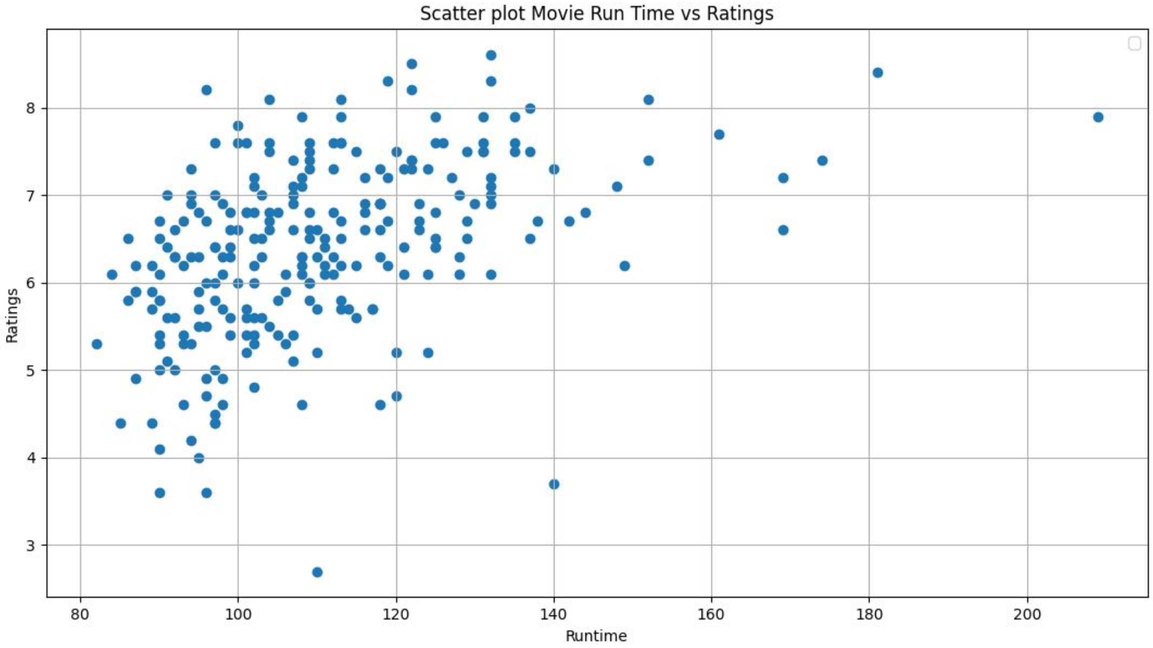
Visualisation and Interpretation of Data:

Distribution of Movie rating and duration:



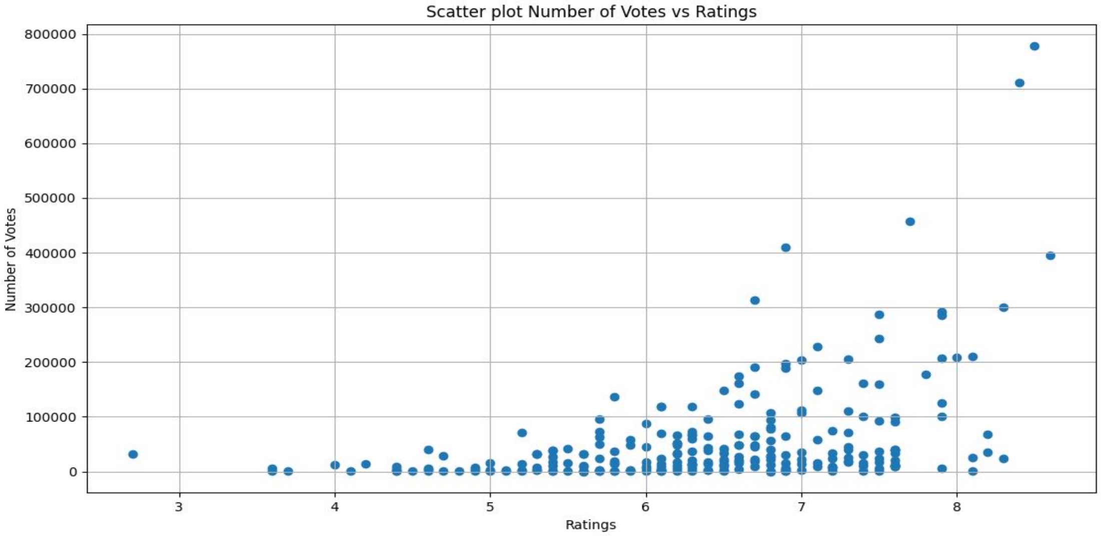
- Ratings: Most of the ratings are between 6/10 to 8/10.
- Duration: Large numbers of movies have duration between 80 mins and 120 mins.

Graphical representation of Movies and duration of films

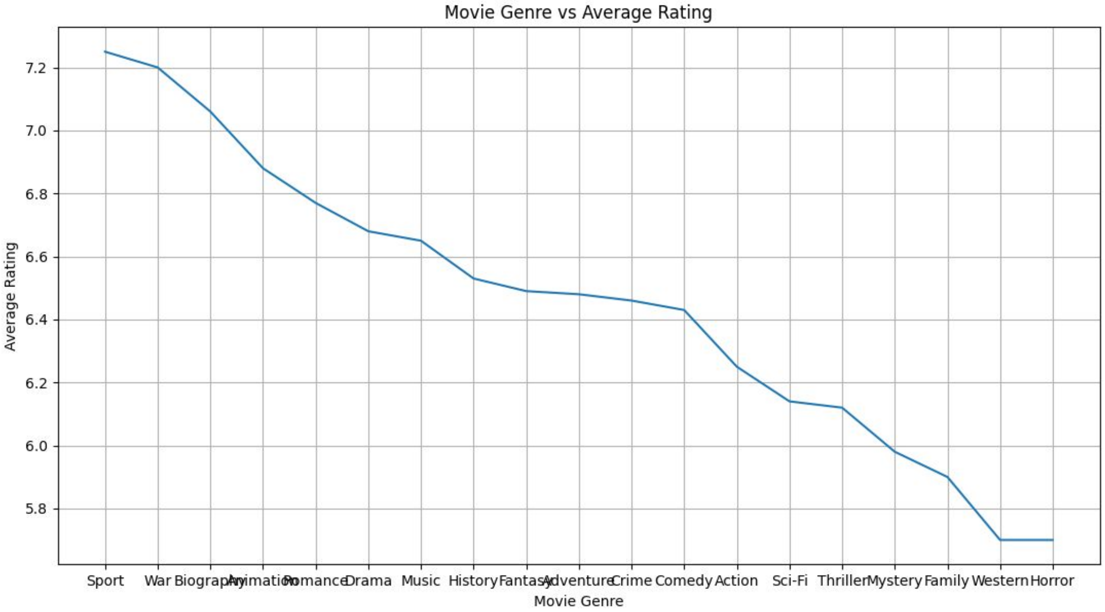


- On this graph we can note that ratings of movies are generally higher for larger runtime movies. Also, for the movies between 60 to 120 minutes the ratings are more concentrated and vary between 4/10 to 8/10.

Graphical representation of the number of votes according to the ratings

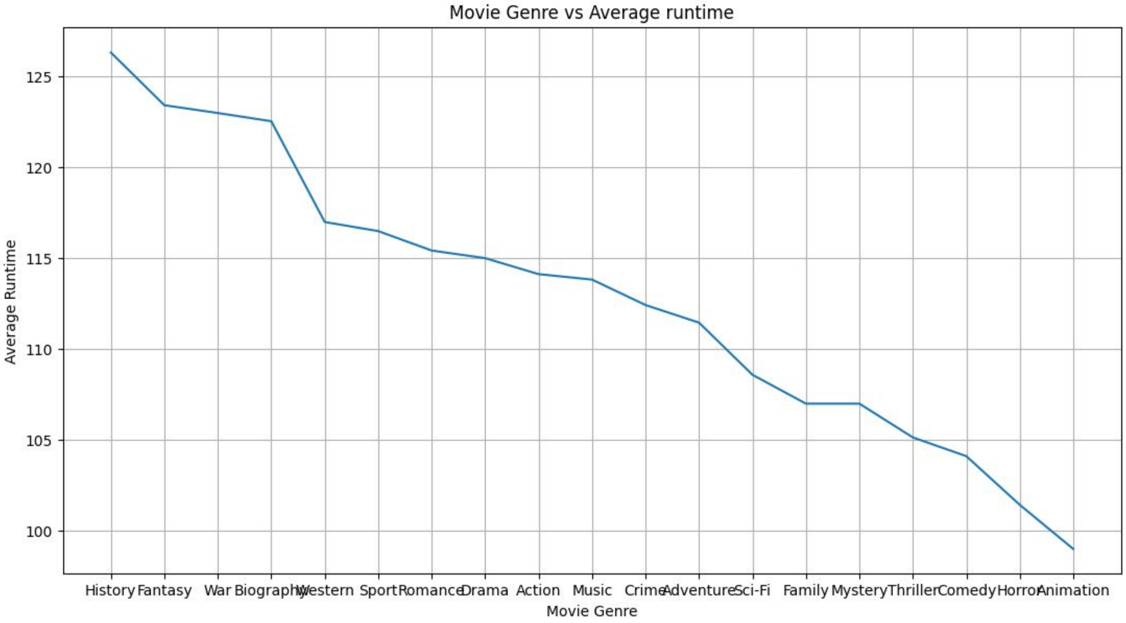


- On this graph we can see that as more people enjoy the movie they tend to vote more and give the movie a higher rating.
  Graphical representation of Average Rating and Movie Genre.



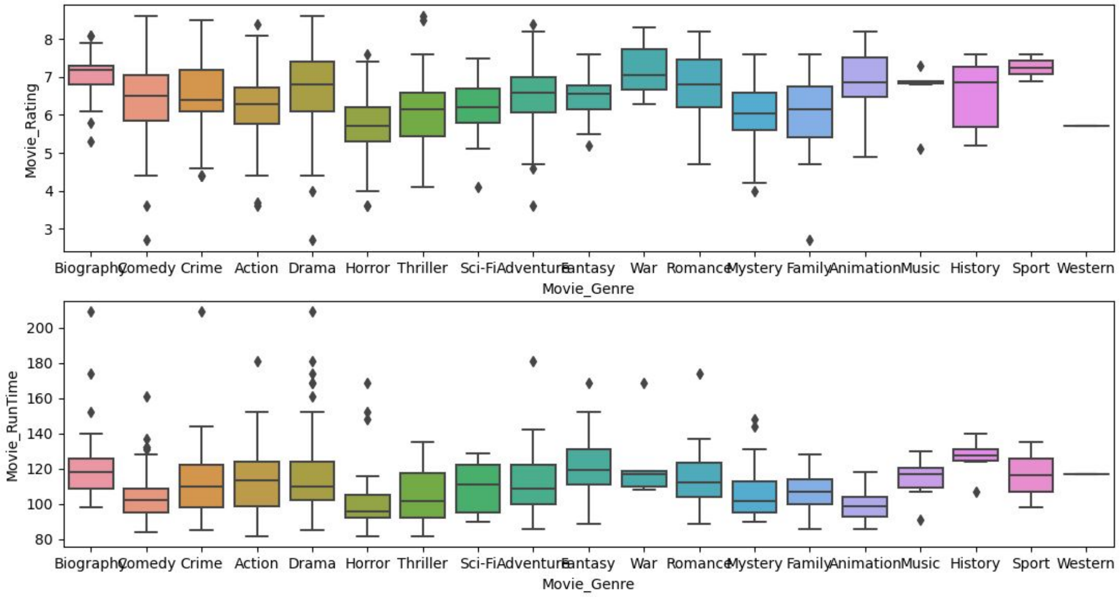
- We can see that the average rating of Sport and War movies in 2019-2020 is highest and horror movies have the lowest average rating.

Graphical representation of Average Runtime and Movie Genre



- On this graph we can note that Average runtime of History and Fantasy movies is the highest whereas Animation and horror movies have lowest average runtime.

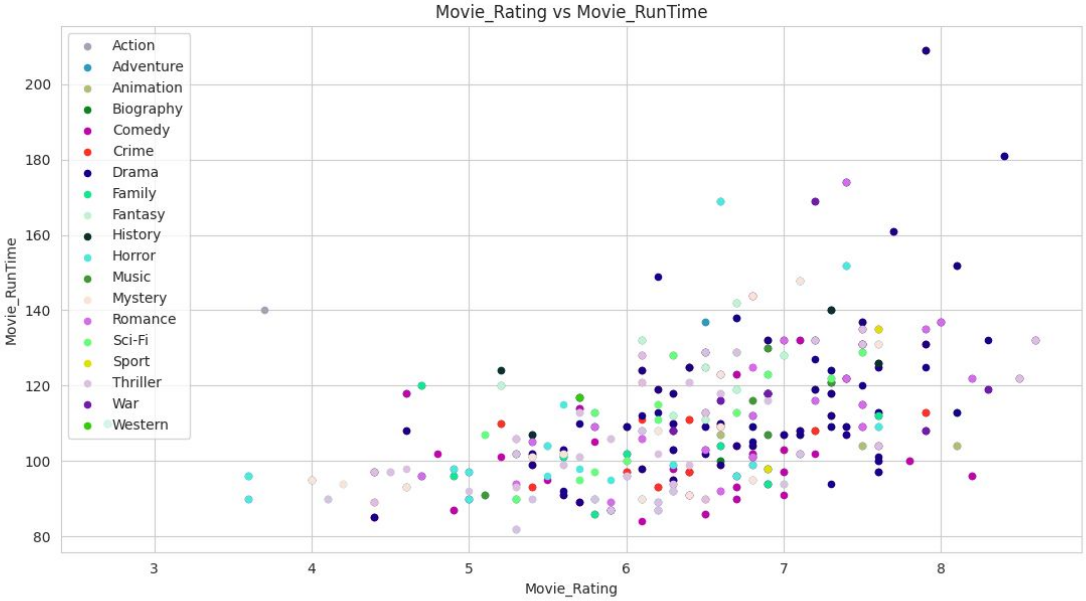
Boxplot of Rating, Runtime depending on the genres of movies in the year 2019-2020:



In these boxplots, one must refer to the median, at the minimum and maximum to have a view of the dispersion of the data around the median.

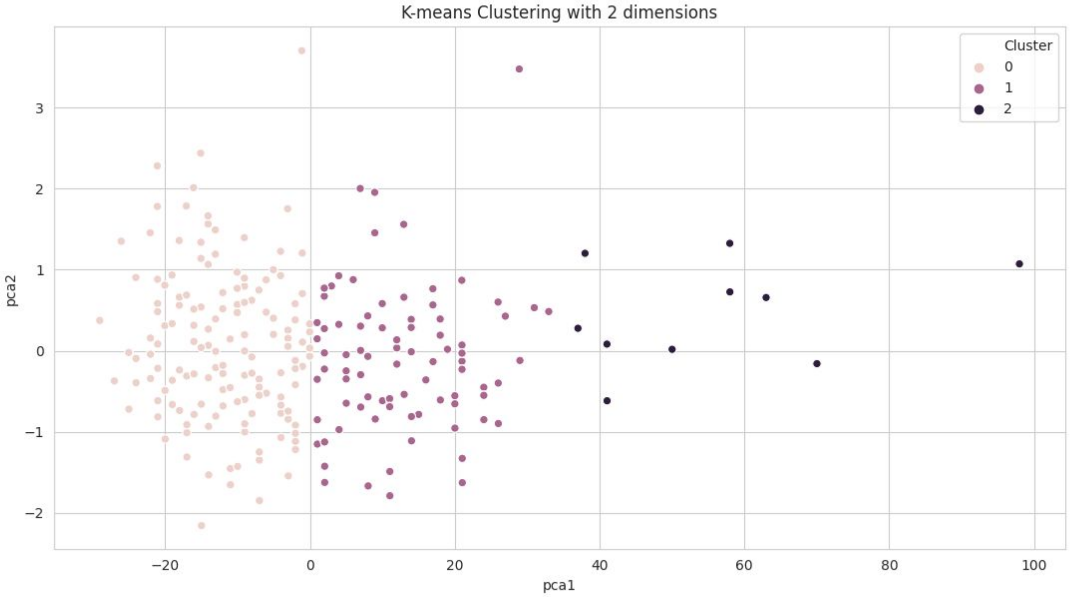
- Movie Ratings: We can see that crime, History and Animation movies are not distributed uniformly around the median.
- Movie runtime: We can note that War, Sci-fi and history movies are not distributed uniformly around the median.

Graphical representation of Movie rating and Movie Runtime with respect to Genre.



- We can see that Movies are distributed uniformly around Movie ratings and Movie Runtime with most of the Movies of Drama Genre having High ratings and high runtime.

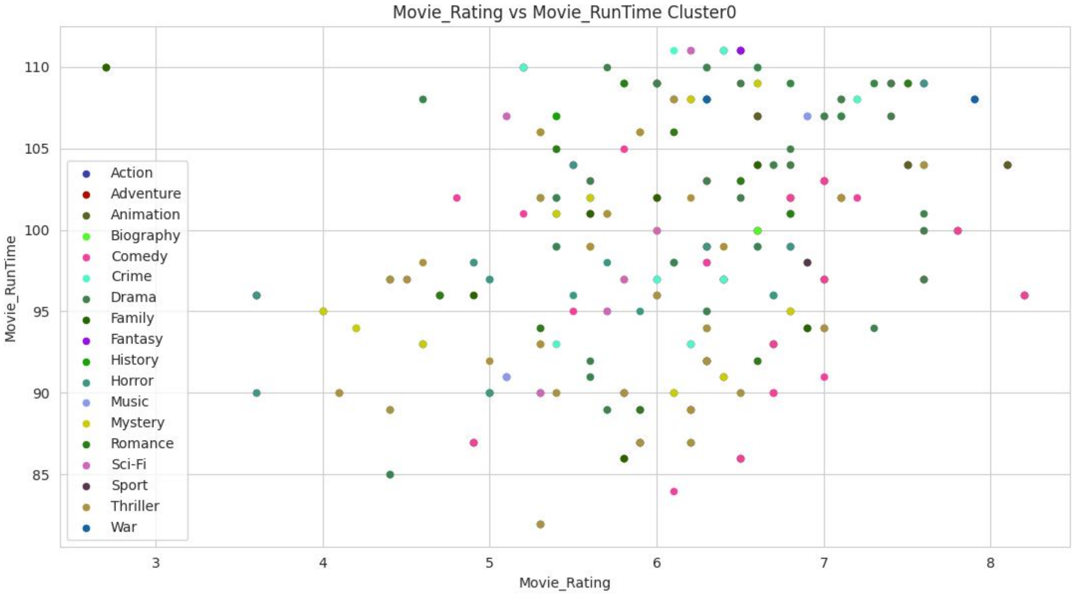
Performing Principal Component analysis and K-means on the data.



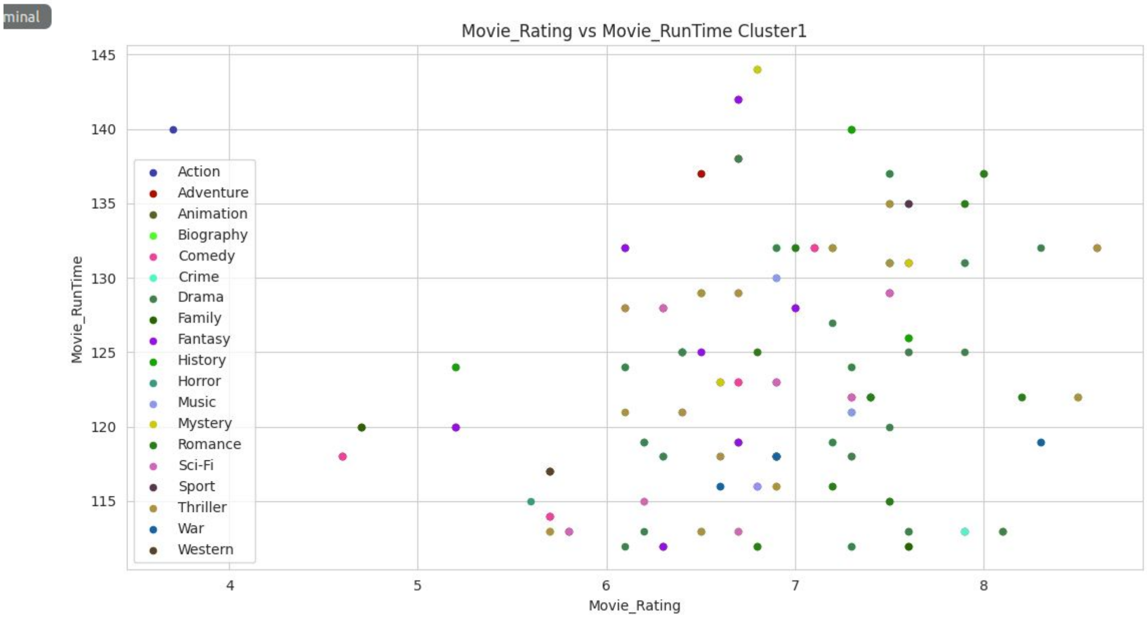
- We can note from the graph that there are Outliers in the data when the clustering is performed Outliers for cluster 1 and cluster 2 can be seen from the PCA graph above.

Graphical representation of Genre present in different cluster

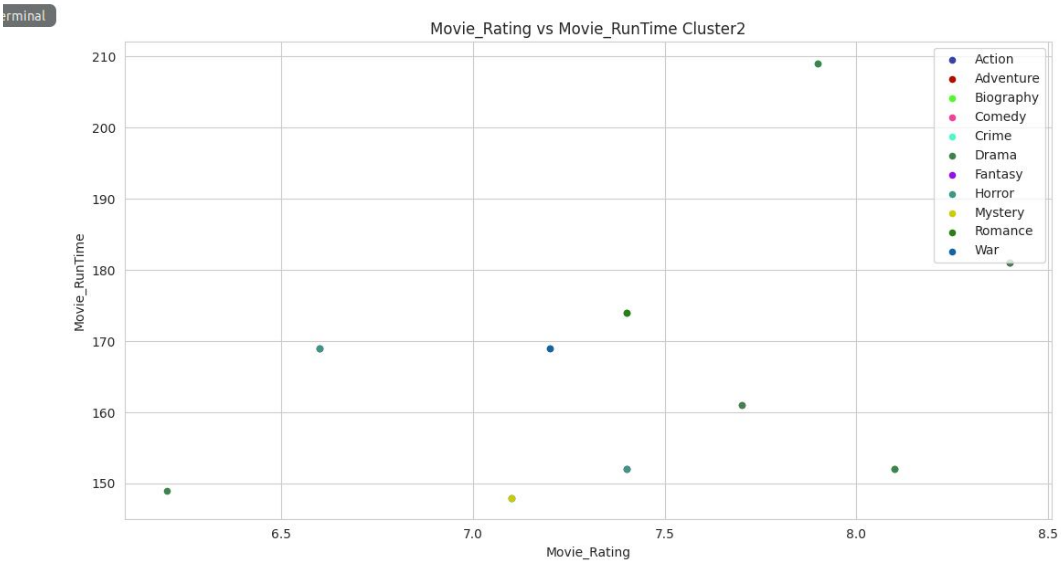
For Cluster 0



- We can see that For cluster 0 Movie has runtime between 85-110 minutes and does not include the Genre Western, Therefore if the new movie comes in with Western Genre we can definitely conclude that it does not belong to Cluster 0 and Does not have runtime between 85-110 minutes.



- We can note that for Cluster 1 the Movie Runtime is between 115-145 minutes. It includes all the genres.



- Runtime of movies in cluster 2 is above 150 minutes and it includes the Clusters Action, Adventure, Biography, Comedy, Crime , Drama, Fantasy, Horror, Mystery, Romance, War.
- Ratings of the movies are above 6 for this cluster.

Conclusion:

The preparation of the data, the modeling of these data, then the visualization of these data with a wide variety of graphs, and finally the interpretation of these graphs made it possible to conduct an analysis for the movies between 2019-2020. We have reached the following conclusions:

- Most of the movies have a runtime between 60 minutes to 120 minutes and ratings between 4/10 to 8/10.
- They tend to vote for a movie if they like the movie. If they don’t like the movie they prefer not to express their opinion.
- Sports,War movies are highest rated and horror movies are lowest rated.
- History movies have the highest runtime and animation movies have the lowest runtime.
- There are very few Mystery and Western movies made.
- From the K-means clustering, we can conclude that if a movie belongs to Genre Western it will have a runtime between 115-145 minutes.
- From the K-means we can infer that If Movie belongs to Genre Action, adventure,crime, drama, fantasy, horror, mystery,Romance, War . It is possible that they may have a higher runtime and ratings above 6.

Future Scope:

- To expand the database by scrapping more data of gross movie Income, movie summary , actors and much more.
- To build the Recommender system for rating the unrated movies on the IMDb website.